



14 - 18 April 2025

HOME

ABOUT

CONTENT

REPORT

CLOSING

BOOTCAMP

DIGITAL SKILL FAIR 38 - DATA SCIENCE

Exploratory Data Analysis - Titanic

Exploratory Data Analysis (EDA) dari dataset Titanic dengan menyediakan analisis data yang mendalam, akurat, dan informatif dengan menggunakan Google Colab dengan bahasa pemrograman Python. Disini kita akan memperlihatkan data-data yang kosong (Null) dan menghapus data duplikat.

by. Haidar Nabiilah Sunu

READ MORE



Data Set of Titanic

Dataset Titanic terdiri dari beberapa kolom penting yang merepresentasikan informasi dasar tiap penumpang. Kolom `survived` menunjukkan status keselamatan penumpang, di mana nilai 1 berarti selamat dan 0 berarti tidak selamat. Kolom `name` berisi nama lengkap penumpang lengkap dengan gelar sosial. Kolom `sex` mencatat jenis kelamin penumpang, yaitu male (laki-laki) atau female (perempuan). Sementara itu, kolom `age` mencerminkan usia penumpang dalam satuan tahun, termasuk pecahan desimal untuk usia bayi. Keempat kolom ini sangat penting dalam analisis awal untuk melihat pola keselamatan berdasarkan usia, jenis kelamin, dan status sosial.

	survived	name	sex	age
0	1	Allen, Miss. Elisabeth Walton	female	29.0000
1	1	Allison, Master. Hudson Trevor	male	0.9167
2	0	Allison, Miss. Helen Loraine	female	2.0000
3	0	Allison, Mr. Hudson Joshua Creighton	male	30.0000
4	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   survived    500 non-null    int64  
 1   name        500 non-null    object  
 2   sex         500 non-null    object  
 3   age         451 non-null    float64 
dtypes: float64(1), int64(1), object(2)
memory usage: 15.8+ KB
```

Catatan Pengamatan :

- Dataset terdiri dari 500 baris dan 4 kolom `survived`, `name`, `sex`, dan `age`.
- Tiga kolom (`survived`, `name`, dan `sex`) tidak memiliki nilai yang hilang (non-null = 500), sedangkan kolom `age` memiliki 451 nilai, berarti ada 49 data yang hilang.
- Tipe data sudah sesuai `survived` bertipe integer (karena biner 0/1), `name` dan `sex` bertipe objek (string/kategori), dan `age` bertipe float karena mungkin ada nilai desimal atau nilai kosong yang menyebabkan konversi otomatis.



Displaying of Data

```
[ ] # Import data
df = pd.read_excel('titanic.xlsx')
df.head() # Import data head [ Menampilkan 5 data teratas dari 'titanic.xlsx' ]
```

```
[ ] # Import data tail [ Menampilkan 5 data terbawah dari 'titanic.xlsx' ]
df.tail()
```

```
[ ] # Import data sample 5 baris data [ diberi secara acak setiap dijalankan ]
df.sample(5)
```

Codingan tersebut adalah tiga cara menampilkan dataset Titanic dengan kode Python yang digunakan di Google Colab, tapi yang pertama yang harus dilakukan adalah mengimport data dari [file titanic.xlsx](#) menggunakan kode `pd.read_excel()`. Berikut penjelasan untuk tiga kode tersebut :

- Pertama kode `df.head()` yang berguna untuk melihat struktur data di bagian atas, yaitu menampilkan 5 baris pertama dari dataset.
- Kode kedua menggunakan `df.tail()` untuk menampilkan 5 baris terakhir dari data, membantu melihat bagian akhir dari dataset.
- Sementara itu, kode ketiga menjalankan `df.sample(5)` yang akan menampilkan 5 baris data secara acak, sehingga setiap kali dijalankan hasilnya bisa berbeda.



14 - 18 April 2025

HOME

ABOUT

CONTENT

REPORT

CLOSING

BOOTCAMP

Statistical Summary Data

Statistical Summary dalam EDA di Google Colab adalah ringkasan statistik deskriptif dari data numerik. Ringkasan ini mencakup nilai `count` (jumlah data), `mean` (rata-rata), `std` (standar deviasi), `min` (nilai minimum), `25%` (kuartil pertama), `50%` (median), `75%` (kuartil ketiga), dan `max` (nilai maksimum). Informasi ini sangat berguna untuk memahami distribusi data, mendekripsi outlier, serta melihat sebaran nilai dari setiap fitur numerik. Dengan melihat hasil statistik ini, analis data dapat menentukan langkah selanjutnya, seperti normalisasi, penghapusan data ekstrem, atau transformasi data sebelum dilakukan pemodelan lebih lanjut.

```
# Membuat kelompok 'categoricals' dan 'numericals'  
categoricals = ['sex'] # Menentukan kolom-kolom kategorikal  
  
# Menentukan kolom-kolom numerikal  
numericals = ['survived', 'name', 'age', ]
```

	survived	age
<code>count</code>	500.000000	451.000000
<code>mean</code>	0.540000	35.917775
<code>std</code>	0.498897	14.766454
<code>min</code>	0.000000	0.666700
<code>25%</code>	0.000000	24.000000
<code>50%</code>	1.000000	35.000000
<code>75%</code>	1.000000	47.000000
<code>max</code>	1.000000	80.000000

Hasil Kelompok Numerikal

sex	
<code>count</code>	500
<code>unique</code>	2
<code>top</code>	male
<code>freq</code>	288

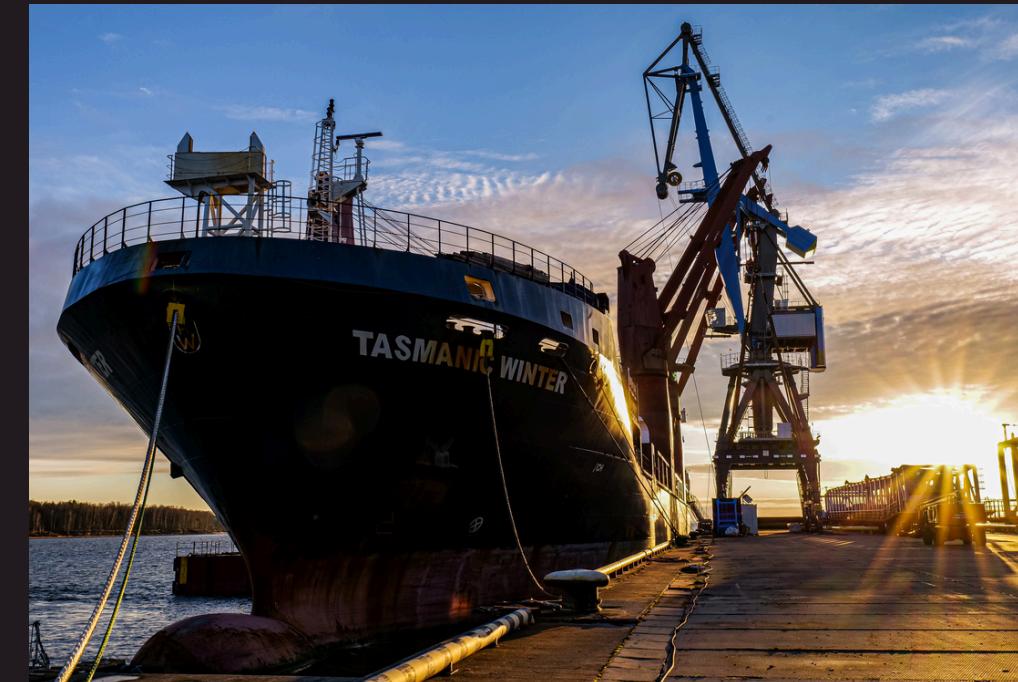
Hasil Kelompok Kategorikal



Data Inspection and Cleaning

Data inspection and cleaning adalah proses awal yang penting dalam analisis data untuk memastikan kualitas dan keakuratan data sebelum digunakan lebih lanjut. Data inspection melibatkan pemeriksaan struktur data, tipe data, nilai yang hilang, duplikasi, dan outlier guna memahami kondisi awal dataset.

Setelah itu, data cleaning dilakukan untuk memperbaiki atau menghapus kesalahan seperti mengisi nilai kosong, menghapus data duplikat, memperbaiki format data, serta menangani data ekstrem atau tidak konsisten. Kedua proses ini bertujuan untuk menghasilkan data yang bersih, valid, dan siap dianalisis agar hasilnya lebih akurat dan dapat diandalkan. Berikut yang akan kita lakukan adalah :



A. DUPLICATE HANDLING

Duplicate handling adalah proses mengidentifikasi dan menghapus data yang tercatat lebih dari satu kali untuk menjaga keakuratan dan konsistensi data.



B. MISSING VALUE HANDLING

Missing value handling adalah cara mengatasi data yang hilang, seperti dengan menghapus atau mengisi nilai kosong pada data agar analisis tetap akurat.



Duplicate Handling

Duplicate handling adalah proses mendeteksi dan menangani data yang tercatat lebih dari satu kali dalam dataset. Duplikasi bisa terjadi karena kesalahan input, penggabungan data dari berbagai sumber, atau proses pengolahan data yang tidak tepat.

Data duplikat dapat menyebabkan hasil analisis menjadi bias atau tidak akurat, karena informasi yang sama dihitung lebih dari sekali. Penanganannya meliputi identifikasi baris yang identik atau sebagian identik (duplikat parsial), lalu memutuskan apakah akan menghapus semua salinan, menyisakan satu data unik, atau menggabungkan data jika diperlukan.

[] duplicates				
	survived	name	sex	age
104	1	Eustis, Miss. Elizabeth Mussey	female	54.0
349	1	Eustis, Miss. Elizabeth Mussey	female	54.0

(kode ini memperlihatkan data-data yang duplikat dan membuat kolom baru untuk mengtahui jumlah duplikat data)



```
[ ] # Hitung frekuensi kemunculan tiap baris duplikat
duplicate_counts = duplicates.groupby(list(df.columns)).size().reset_index(name='jumlah_duplikat')

# Urutkan berdasarkan jumlah duplikat
sorted_duplicates = duplicate_counts.sort_values(by='jumlah_duplikat', ascending=False)

# Tampilkan hasil
print("Baris duplikat yang sudah diurutkan berdasarkan jumlah kemunculannya:")
sorted_duplicates

[ ] Baris duplikat yang sudah diurutkan berdasarkan jumlah kemunculannya:
[ ] survived name sex age jumlah_duplikat
[ ] 0 1 Eustis, Miss. Elizabeth Mussey female 54.0 2
```



Missing Value Handling

Missing value handling adalah proses menangani data yang hilang agar tidak memengaruhi hasil analisis. Caranya bisa dengan menghapus baris atau kolom yang kosong, atau mengisi nilai yang hilang menggunakan metode seperti mean, median, modus, atau algoritma prediktif, tergantung pada jenis dan konteks data. Tujuan utamanya adalah menjaga kualitas dan keakuratan data.

```
[ ] # Menghitung total baris dalam DataFrame
total_rows = len(df)

# Loop untuk setiap kolom di DataFrame
for column in df.columns:
    # Hitung jumlah nilai yang hilang (NaN) di kolom
    missing_count = df[column].isna().sum()

    # Hitung persentase nilai yang hilang terhadap total baris
    missing_percentage = (missing_count / total_rows) * 100

    # Tampilkan jumlah dan persentase nilai yang hilang
    print(f"Column '{column}' Has {missing_count} missing values ({missing_percentage:.2f}%)")

→ Column 'survived' Has 0 missing values (0.00%)
Column 'name' Has 0 missing values (0.00%)
Column 'sex' Has 0 missing values (0.00%)
Column 'age' Has 49 missing values (9.82%)
```

Hasil Output :

- Menghitung total baris pada DataFrame.
- Mengecek jumlah nilai hilang (NaN) di setiap kolom.
- Menghitung persentase nilai hilang terhadap total baris.
- Menampilkan jumlah dan persentase missing values per kolom.
- Hasilnya hanya kolom 'age' yang memiliki missing values (9.82%).





14 - 18 April 2025

HOME

ABOUT

CONTENT

REPORT

CLOSING

BOOTCAMP

The Report of Analysis



Kesimpulannya, analisis awal pada dataset Titanic menunjukkan pentingnya proses data inspection dan cleaning dalam memastikan kualitas data sebelum digunakan untuk analisis lanjutan. Dengan mengidentifikasi adanya data duplikat dan nilai yang hilang, terutama pada kolom usia (age), langkah-langkah penanganan seperti penghapusan atau pengisian nilai kosong dapat dilakukan untuk menjaga keakuratan hasil. Proses ini membantu menciptakan data yang bersih dan konsisten, sehingga mendukung analisis yang lebih valid dan dapat diandalkan.

Let's Connect :



[GitHub - Haidar Nabiilah Sunu](#)



[LinkedIn - Haidar Nabiilah Sunu](#)



14 - 18 April 2025

HOME

ABOUT

CONTENT

REPORT

CLOSING

BOOTCAMP

Thank You for Your Attention ... !



SMK TELKOM PURWOKERTO



BOOTCAMP @DIBIMBING.ID



PROJEK PORTOFOLIO